Geocoding in Cancer Research
A Review

Gerard Rushton, PhD, Marc P. Armstrong, PhD, Josephine Gittler, JD, Barry R. Greene, PhD, Claire E. Pavlik, PhD, Michele M. West, PhD, Dale L. Zimmerman, PhD

Abstract: There is now widespread agreement that geographic identifiers (geocodes) should be assigned to cancer records, but little agreement on their form and how they should be assigned, reported, and used. This paper reviews geocoding practice in relation to major purposes and discusses methods to improve the accuracy of geocoded cancer data. Differences in geocoding methods and materials introduce errors of commission and omission into geocoded data. A common source of error comes from the practice of using digital boundary files of dubious quality to place addresses into areas of interest. Geocoded data are linked to demographic, environmental, and health services data, and each data type has unique accuracy considerations. In health services applications, the accuracy of distances computed from geocodes can differ markedly. Privacy and confidentiality issues are important in the use and release of geocoded cancer data. When masking methods are used for disclosure limitation purposes, statistical methods must be adjusted for the locational uncertainty of geocoded data. We conclude that selection of one particular type of geographic area as the geocode may unnecessarily constrain future work. Therefore, the longitude and latitude of each case is the superior basic geocode; all other geocodes of interest can be constructed from this basic identifier.
(Am J Prev Med 2006;30(2S):S16–S24)

Introduction

Geocoding is the practice of assigning a geographic identifier to a computer record that lacks it, thereby tying information to geographic space. Although one of the first applications of geocoding was made for health data,[1,2] modern geocoding practices were more fully developed outside of health. More recently, however, a number of groups have advocated using geocoded health data in geographic information systems (GIS) as an essential component of health systems analysis.[3–6] There are a number of published introductions to geocoding materials and methods.[7,8] Missing from the geocoding literature, however, is a consideration of how the characteristics of geocodes should be matched to their intended uses. This article reviews geocoding practices in relation to how they are used in cancer research and discusses methods to improve the accuracy of geocoded cancer data.

From the Department of Geography (Rushton, Armstrong, Pavlik), Department of Epidemiology (West), Center for Health Policy and Research (Rushton, Greene), Department of Health Management and Policy and Center for Health Policy and Research (Greene, Pavlik), College of Law (Gittler), and Department of Statistics and Actuarial Science (Zimmerman), University of Iowa, Iowa City, Iowa

Address correspondence and reprint requests to: Gerard Rushton, PhD, Department of Geography, The University of Iowa, 316 Jessup Hall, Iowa City IA 52242. E-mail: gerard-rushton@uiowa.edu

Traditionally, geocoding involved assigning census or other geographic codes to health records. In the U.S., Federal Information Processing Standards (FIPS) (see Appendix for explanations of acronyms) codes[9] have standardized the identification of geographic entities and their geographic coordinates. Although Howe[10] has discussed the benefit of using fine geographic detail (individual addresses) to analyze cancer registry data, general use of more precise geocodes is more recent. Analyses that require this level of detail involve, for example, establishing relationships between cancer and environmental contaminants when interest is focused on the distance between a pollution source and exposure to it. When short distances are associated with health effects, the geocode must have a positional accuracy that is sufficient to resolve whether such effects are present. Where the focus is on the health effects of environmental contamination, many environmental conditions have spatial patterns that do not correspond with existing political or administrative areas. For example, it might be necessary to determine whether the cancer incidence rate for people whose water supply is drawn from a particular aquifer differs from the rate for those living elsewhere. Examples of questions about the relationship between locations of cancer and environmental factors have been examined in previous studies.[11–17] The health disparities literature requires the measurement of relationships between cancer incidence or mortality rates and the

0749-3797/06/$–see front matter
doi:10.1016/j.amepre.2005.09.011

socioeconomic characteristics of people. Such questions require cancer geocodes that link to the geocodes of demographic data. Other research has examined cancer stage at diagnosis in relation to residence within particular hospital catchment areas.[18,19]

Methods

There are three major methods of geocoding. Each has a particular set of supporting materials and characteristic errors. The first method assigns an observation (a cancer record) to a geographic unit. Location within the unit is not specified, but ecological analyses can be carried out using data associated with the geographic unit. The second and third methods, interpolation and parcel matching, attach a point coordinate value to a record. The interpolated method (described in the next section) estimates a coordinate based on a proportional distance between the address on a record and the address range for a street segment. With the parcel method, each parcel has an address and a specific coordinate is assigned to it with the coordinates based on either its centroid or the location of the property's major structure, usually to identify the location of the major structure. The parcel method is more accurate and is now becoming widespread given the development of parcel level databases by many U.S. cities and counties. In the following section, interpolated and parcel address-matching are used to illustrate geocoding methods, materials, and error characteristics.

Interpolated and Parcel Matching Methods

Address matching links a postal address to a geographic base file that contains geographic units and coordinates. In the U.S., TIGER (Topologically Integrated Geographic Encoding and Referencing) files or commercial derivatives are often used for address matching. TIGER files[20,21] are structured such that for each street segment, usually from intersection to intersection, a range of addresses is coded for each side of the street. These files also contain the IDs of the Census geographic entities to which the street segment in question belongs. These IDs are used by the Census to correctly assign each address to Census geographic areas in the data tabulation process. To address match a health record, the correct street and street segment are determined first; ancillary (e.g., ZIP code) information is often used to localize searches. Next, the address for the health record is made proportional to the range of addresses for that street segment. For example, if a segment has addresses from 101 to 199, and the health record has a value of 149, the address is calculated to be halfway along the segment. Finally, the same proportion is applied to the difference in the coordinate values for the endpoints of the segment, and the coordinate values for the address are interpolated between the endpoints.

In the parcel matching method, an address is linked to a geographic file that contains specific, known, addresses—sometimes called a master address list. These geographic files may come from several sources. E911 files, for example, link telephone numbers to locations so that emergency responders can find a caller's physical location based on telephone numbers alone. In other cases, land parcel files delimit the boundaries of legally defined land parcels. In rural areas, however, these parcels may be large and contain many structures, so that a residence location may be different from the centroid of the parcel, a common coordinate location used in parcel data. Normally, however, parcel files provide more accurate locations of addresses than those obtained using interpolated locations from street centerline files.[22]

Main Components of Geocoding Error

The proportion of addresses correctly matched (hit rate) increases if efforts are made to increase the quality of the address file and the geographic base file. The geographic file may have inaccurate geometry and there are many consequences of geometrical errors. Several commercial firms offer street centerline files derived from TIGER-Line files in which their geometrical accuracy and attribute contents have been enhanced.[23,24] Other errors can also occur, such as missing streets or streets that do not connect in reality but are encoded as such. These types of errors occur most frequently in areas experiencing new development, and the need to eliminate them is a key reason for the emergence of a private sector address-matching industry. Errors in geographic files can be reduced through fieldwork, using global positioning systems (GPS) observations, and by using accurate, updated local maps.

Address-matching errors occur even when a highly accurate geographic base file is available. For example, in some cases, round numbers (100s) are used to specify an address range, even though the actual addresses may run into only the first half of that range. Because interpolation is based on an address range that is too large compared to the actual address range, interpolated coordinate values will be displaced toward the low end of the street segment. Misspellings and incorrect address formats in the health records to be geocoded are also problematic. In some cases, approximations are made using SoundEx algorithms, but this practice does not ensure correct results. Consequently, it is a good practice to ensure that a standardized set of abbreviations is used, along with correct specification of street prefixes and suffixes (e.g., 123 Northwest Boulevard Court and 123 NW BLVD CT). Standards exist for street names and addresses in the U.S. and the U.S. Postal Service provides assistance in meeting these standards.[25–27] In some cases, however, residence locations cannot be located because a post office box is used. Then a rule must be established to assign coordinates to these addresses.[28,29]

Commercial geocoding software permits users to define the conditions under which an address is considered matched by allowing them to set a match confirmation threshold score. Such scores are commonly additive weights on matches made to components of the address. A perfect score means that exact matches were made to all components of the address.[30] Given the deficiencies of geographic base files and the common errors in street addresses, confirmation match thresholds permit users to increase match rates by permitting less than exact matches. For a given real-world set of addresses, there is thus a trade-off: increasing the match rate by

lowering the threshold score results in a decrease in accuracy and therefore geocoding quality.

Assessing the Quality of Geocodes

For many uses of geocoded cancer records, addresses are linked with census areas to determine characteristics of the area in which an address is located. Accuracy, therefore, is based on the assignment of an address to its correct census unit. Because TIGER files were originally developed to assign addresses to census tabulation areas, this linkage is best done by using the same methods and materials used by the U.S. Census. Unfortunately, some commercial geocoding software programs do not use this approach. Instead, they use GIS "point-in-polygon" procedures to link geocoded address coordinates to census areas. This approach relies on the accuracy of the geometry of the census area boundary files. These files are notoriously inaccurate, however, and the U.S. Census itself indicates that "cartographic boundary files should not be used for geocoding."[31] It is well known that errors in spatial assignment to any administratively defined areas can be large when the absolute location of points with respect to boundaries is the basis for the spatial assignment.

A comparison of two commercial geocoding products, one of which uses the point-in-polygon method and a second that uses the census area IDs from TIGER files, found that 28% of addresses were not assigned the same census block and that 51% of addresses with different blocks did not have the same census tracts.[30] The conclusion of Yang et al.[30] that the best strategy to improve geocoding would be to use "further enhanced street reference data for accurate geographic location information" only reinforces the false impression that more precision in the geographic base files is the key to improving the accuracy of the census area IDs assigned. Instead, the key is to use the superior method: census area IDs should be assigned using a look-up table that links the address to the street segment in the TIGER file that contains the census area IDs of that street segment. They should not be based on geometric point-in-polygon procedures.

Issues that Arise Using Geocoded Cancer Data

Increasingly, geocoded cancer data are used to link to socioeconomic, demographic, or environmental data as well as to health service locations for screening or therapy. Cancer geocodes must be compatible with these other geocoded data. Geographic reference coordinates may be specified as longitude and latitude coordinates or as one of the standard systems that express latitude and longitude as projected coordinates.[40] Various geocoding issues arise, depending on the type of linkages made.

Linking Geocoded Cancer Data to Demographic Data

When cancer records are linked to demographic data, incidence rates can be computed only when both cancer incidences (numerators) and people at risk (denominators) are available for the same geographic areas.[41] Within a cancer registry, the computation of such rates for flexibly defined areas is often constrained more by the lack of appropriate population data than the availability of cancer data. For users who rely on access to publicly accessible cancer data, confidentiality restrictions limit linkages to the areas for which the cancer data is made available. Users of geocoded cancer data often seek to know the smallest geographic unit for which both cancer data and detailed, age/gender, population data exist. The answer in the U.S. is usually the census block group. Census tracts for which detailed population data are also available are approximately four times larger than census block groups. The advantage of census tracts is that they are similar in population size and, in many but not all cases, in socioeconomic characteristics.[42]

There has been considerable discussion about problems in using five digit ZIP codes as geocodes.[4,39,43] The boundaries of such areas change through time and the changes are not well-documented. Since 1990, the U.S. Census has not tabulated its data by ZIP code and, starting with the 2000 census, has instead published tabulations for ZCTAs (ZIP Code Tabulation Areas). ZCTAs are aggregations of Census blocks that most closely correspond with ZIP code areas. These areas are often seriously misaligned as compared to the areas served by the ZIP code.[44] The ZCTA demographic data presumably are poorer representations of actual ZIP code demographic characteristics than the 1990 ZIP code tabulations were.[4,45] Whether knowing the location of the boundaries of ZCTAs compensates for this loss of data accuracy is for the data user to decide. ZCTAs are only likely to be useful in cancer surveillance and research if cancer records are geocoded to census blocks because cancer data could then be aggregated correctly to ZCTA for which population characteristics are known. We are unaware of any example yet of any researcher or cancer registry that has done this. The Health Resources Services Administration (HRSA) uses ZCTAs to display Medicare data although one has to presume that their analyses of Medicare data are based on the corresponding ZIP codes in their data files. Their website states, "the PCSAs were developed by aggregating ZIP Code Tabulation Areas," although this is surely impossible because the geocode on the record of each beneficiary is their ZIP code, not their ZCTA. In this example, the analyses of Medicare data use the ZIP code, the displays use ZCTA boundaries and any relating of the Medicare data to socioeconomic or demographic data is for the spatially misaligned ZIP codes

and ZCTAs. This is a path that cancer registries should avoid taking and they can do so by geocoding their records to the Census block and then aggregating them to the ZCTA with the result that cancer records and socio-economic and demographic data will generally be for the identical spatial units. In practice, however, as ZIP code boundaries change between census dates, some spatial misalignment would still occur. When users of geocoded cancer data cannot obtain tabulated population data for the geographic area of their interest, a possible solution is to obtain population data through the process of areal interpolation.[46–49]

Linking Geocoded Cancer Data to Environmental Data

Although there is an emerging consensus that census tracts and block groups are valuable geocodes for linking cancer data to socioeconomic characteristics, no such consensus exists for environmental data. Vine et al.[50] provide an introduction to the subject. Environmental data commonly exists as remote sensing imagery and are, therefore, more naturally represented in raster rather than the vector data form generally used to record demographic data.[7,51] Consequently, geocodes are often spatially misaligned and it is difficult to measure relationships between health outcomes and environmental measures. Recent studies of agricultural pesticide use and childhood cancer in California illustrate this problem.[17,52] In these studies pesticide use locations were coded to approximately one-square-mile areas and population data were available by census block group. The spatial misalignment of these areas, in addition to their large size, made it difficult to accurately compute the relationship between pesticide application and cancer incidence rates.

Linking Geocoded Cancer Data to Health Services Data

Geocodes are used to estimate distances between cancer cases and health services. There are three ways to find distances from geocodes.[7,53–55] If the geocodes are areas such as census tracts, distances can be computed from their centroids.[7,53,56] These distances are rough estimates but they may be adequate for some purposes.[56,57] Errors occur from two sources: first, the estimated distances do not take into account the road network and possible routes that a patient may select in using a health facility; second, error occurs when the locations of many people are assumed to occur at the centroid of the area. This is known as spatial aggregation error.[58,59] If the geocodes are coordinates located on a road network and if the road network has been coded with distances between all nodes, then shortest distances from the known location to all other nodes on the network can be computed using a shortest path algorithm,[60,61] although there is widespread appreciation of the fact that people will often not use their nearest health facility.[61] Finally, there are now advanced spatial analysis methods that evaluate the accessibility of health facilities within the context of the spatial–temporal constraints that individuals face.[62–64]

Privacy and Confidentiality Issues

Geocoding cancer records raises a host of public policy, legal, and technical issues that are related to the privacy of personal health information. These illustrate the growing concerns about the threats to individual privacy posed by the dramatic advances in information technology.[80–82] In many western countries, concerns include how new geospatial technologies may be used to infringe on people's privacy.[83–89] Additional privacy concerns arise as a consequence of being able to accurately represent the location of individuals using geocoding and GPS, as well as the ability to link disparate data sources. Geographic identifiers support such linkages because data are easily combined when common identifiers such as names, social security numbers, phone numbers, or driver license numbers or home or work addresses are present in different databases. This is true of health records as well as many other data sources.

When data are integrated from various sources, the widespread availability of GIS software has led to the common practice of creating maps of the results. For example, a point map of cancer incidences can be created in much the same way as Snow's famous "Cholera Map." If such maps accurately depict locations, they can be used to recover individual-level information such as an address. This process, referred to as "inverse geocoding,"[83] first determines an address from the location of each map symbol, then uses it to forge a link to other data sources.

Personal health information in medical records is widely regarded[90,91] as private, deserving of protection from unauthorized and unwarranted invasion. Thus, numerous laws afford protection to the privacy of this type of information.[92,93] Nevertheless, personal health information is not entitled to absolute privacy protection. Rather, a patient's interest in the privacy of personal health information must be balanced against societal interests in protecting and promoting public health.[94]

The initial issue that must be addressed in connection with the geocoding of cancer records is whether and under what circumstances they may be used for geocoding. In 2002, the U.S. Department of Health and Human Services issued regulations, known as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[95] This rule generally provides that a defined set of "covered entities" cannot disclose personal health information without authorization from the individuals involved. However, the rule creates an exception permitting

disclosure of personal health information to public health authorities for public health purposes without such authorization. The Centers for Disease Control and Prevention (CDC)[96] and several other agencies have interpreted this exception as permitting healthcare providers to disclose cancer cases to a state cancer registry pursuant to state law. However, these public health authorities must take steps to ensure that specific individuals cannot be identified. Some state statutes and administrative regulations specifically require cancer registries to take steps to safeguard the security and confidentiality of registry data.[97] Geocoding methods are currently being developed to be consistent with such requirements. For example, Krieger et al.[19] described a procedure designed to access individual-level health data held by a public health agency without compromising the privacy of this data. However, according to the North American Association of Central Cancer Registries (NAACCR)[98] "a major challenge is developing mechanisms to appropriately protect, secure and release data while protecting patient confidentiality in a standard fashion so that each cancer registry does not have to develop them on their own."

Cancer registries distinguish between publicly released data and data made available to researchers. Publicly released cancer data are released in a form that is directed at protecting the identity of individuals. This is commonly done by ensuring that the release is of aggregated data from populated areas with a minimum number of persons as well as a minimum number of cases in any category released.[77] Finer geographic detail is sometimes provided to researchers through the mechanism of data-use agreements.[77,99] When institutions, such as departments of public health, have access to individual health records but are bound by law to protect the confidentiality of their data, there are methods to geocode these records consistent with privacy requirements—see discussion below of "masking."

Statistical Methods, Geocoding Errors, and Disclosure Limitation

Geocoding errors generally have an adverse effect on statistical analyses of cancer data. For instance, the power to detect a cancer cluster over a region decreases as the magnitudes of geocoding errors increase, and is severely degraded if these magnitudes are on the same order as the cluster size itself.[100,101] Similarly, the ability to create accurate fine-level cancer maps or to relate cancer incidence to a geographically varying risk factor is severely limited if geocoding errors are commensurate with the fine-level scale of the map or the scale at which the risk factor operates.

Historically, locational uncertainty due to geocoding errors has been ignored in statistical analyses of cancer data. Recently, however, methods have been developed to account for these errors. Arbia et al.,[102] for example, proposed a "corruption model" that quantifies the effect of locational uncertainty on output map quality. Another approach[103,104] uses measurement error models to incorporate location errors into the estimation of spatial dependence and spatial prediction. Conclusions drawn from these kinds of analyses are generally more appropriate than those drawn from analyses that ignore geocoding errors.

Although considerations such as flexibility, accuracy, and statistical power argue for geocoding cancer data at the finest spatial resolution possible, protection of the confidentiality of individuals may require that accurate geocodes be modified, or "masked," before release of the data to researchers and others. Several geographic masking methods exist,[78] including affine transformations, aggregation (both point and areal), and random perturbations—described as "jittered" in a recent study.[13] The goal of these masks is to modify the geographic information sufficiently to prevent disclosure of individual identities, while retaining enough spatial accuracy for geographic trends, clusters, or other patterns to be detected.[105]

Every geographic mask results in some information loss; however, some or all of the information lost by masking may not be needed to answer a given question. For example, a "nearest-neighbor mask," in which only the distances of each cancer event to the nearest other cancer event are provided to the user, is sufficient for many clustering tests. However, users of these masked data would be precluded from identifying the locations of clusters. Thus, it is clear that a suitable mask for general use should be robust enough to support the investigation of a variety of research/control questions.

Overall, random perturbation seems to be the most robust mask, and not coincidentally it has received the most recent attention.[78,79] It is possible for individual registries to conduct studies of the effect of increasing perturbation on cluster detection and disclosure risk with their cancer datasets to determine a threshold level of perturbation at which there is an acceptable tradeoff between analytic power and disclosure risk. Questions remain as to how much information about a specific perturbation model, i.e., "mask metadata," should be disclosed to subsequent users, as well as the extent to which the combination of multiple masked releases of the data by colluding data users may compromise confidentiality.

Recommendations for Geocoding Practices in Cancer Registries

The North American Association of Central Cancer Registries (NAACCR)[98] developed a handbook of basic practices on the use of GIS with cancer registry data.[24] To improve the process of obtaining accurate addresses, their workgroup recommended that reporters of cancer information be supplied in the field with

Table 1. Questions on geocoding quality

Quality criterion	Reference	Conclusion
How complete are census tract assignments?	Boscoe (2002)[32] McElroy (2003)[33]	Pre-processing of addresses can substantially increase the percent of addresses successfully assigned to census tracts.
Which census areas have the highest proportion of records both geocoded and linked to census-defined areas?	Krieger (2003)[34]	Census tracts
What is the effect of incomplete geocoding on the accuracy of rate maps?	Ratcliffe (2004)[35]	For mapping crime rates, 85% is a minimum acceptable geocoding hit rate.
What is the positional accuracy of rural and urban address-matched geocodes?	Bonner (2003)[36] Cayo (2003)[22] Dearwent (2001)[37]	Distances between geocoded locations and true locations are greater for rural than urban locations.
Are characteristics of people and geocode quality related?	Gregorio (1999)[38]	Women in Connecticut diagnosed with breast cancer were more likely to be successfully geocoded if they were African-American, or lived in urban areas, or in census tracts with low median family incomes.
What are the disadvantages of using five digit ZIP codes as a principal geocode?	Krieger (2002)[4]	They are too large in population, have boundaries that change and, particularly since 2000, they do not have reliable demographic data.
What are the advantages of using five digit ZIP codes as a principal geocode?	Wing and Reynolds (1988)[39]	Health data are widely available for them.
What are the advantages of using finer geographic scales for geocodes?	Boscoe (2002)[32]	Important where a local point source environmental exposure is being investigated.

maps and other information to facilitate the recording and checking of patient addresses when cases are first abstracted. Three general steps should then be taken at the central registry to ensure the maximum value of case level data for geographical analyses: (1) address checking, (2) address correction, and (3) address geocoding. One registry has developed procedures to resolve address-at-diagnosis discrepancies when it receives multiple reports for the same person.[106] Several of the workgroup's recommendations relating to GIS were designed to improve the geocoding process:

1. Latitude and longitude coordinates should be stated in decimal degrees and the North American Datum 1983 (NAD83) should be used.
2. The provisional address data content standard of the Federal Geographic Data Committee[107] should be adopted.
3. A written policy on confidentiality and disclosure rules should be developed and NAACCR should participate in Federal policy and standard setting.
4. Cancer registries need trained staff in the GIS area and should commit time and resources to continuing education.

To these we add that registries should continue to build on and extend the census tract certainty code system recommended by NAACCR to other levels of geography, e.g., the census block group, and should evaluate and report on methods used to verify accuracy of geocodes.[5] Questions relating to aspects of the quality of geocodes are listed in Table 1 with references to work that evaluates each criterion. Work that recommends appropriate geocoding approaches for specific geocoding purposes are listed in Table 2.

Conclusion

Because the key test of the adequacy of any cancer geocode is whether it meets the standard of fitness for its intended uses, no single "best choice" geocode can be determined for all choices. It is the nature of questions asked that must drive the geocode choice for cancer records.

Rather than judge the geocoding requirements for cancer data by the current practices of public health and research communities, requirements should be related to the larger goals of cancer surveillance and control and "best geocoding practice" standards for achieving these should be developed. Unless these goals are given priority, the availability to researchers of geocoded data that is restricted to large administrative areas, such as counties,[41,108] will continue to inhibit the development of better spatial analytic methods to support not only spatial epidemiology, but also more effective and targeted cancer prevention and control.

Strategies developed for geocoding cancer cases should be developed with an eye toward systematic analysis of key relationships between cases and socioeconomic and environmental conditions. This view calls for geocoding that enables assignment of individ-

Table 2. Geocoding recommendations in relation to geocoding purpose

Purpose of the geocode	Recommended geocoding approach	Comments and relevant literature
For a given address, find correct census area such as census tract or block group	Use TIGER Line file that matches the census date. Do not use a "point-in-polygon" GIS procedure. Do not use census boundary files; especially do not use cartographic boundary files.	Krieger (2003)[34]
Find the most accurate latitude and longitude coordinates for an address.	From most accurate to least accurate: GPS E911 Address-match to enhanced street reference file Address-match to latest version of TIGER file	Pre-process addresses to correct format and spellings. Avoid use a low address-match score where this is an option. Examine nonmatches carefully and correct them using best available information
Make different spatial coverages compatible when they have come from a variety of sources.	Decide on a common coordinate projection system. A common choice is UTM	Some GIS do not include required functions to do this. In such cases, the data conversion should be outsourced.
Find the smallest area for which both cancer data and age-gender, race, defined population data I are available.	The Census block group is the best area for this purpose. Census blocks have such data but the swapping of person characteristics to prevent disclosure substantially diminishes its accuracy	Spatial disaggregation methods: Gotway (2002),[65] Flowerdew (1989, 1991, 1994)[48,66,67]
Find population data for area that is not compatible with any census defined area	Select method of spatial disaggregation using polygon overlay or spatial smoothing More accurate methods, generally, are the "intelligent" methods Select raster population data source where available	Above references plus Markoff (1973),[68] Goodchild (1980),[69] Lam (1983),[70] Goodchild (1993),[71] Sadahiro (1999),[72] Langford (1991),[73] Dobson (2000)[46]
Find a smoothed cancer rate based on small-area cancer and population data	Control the "spatial support area." Aggregate cancer and population data for a defined filter area (block kriging) and then compute rate. Do not compute small-area rates then weight them as in Finnish Cancer Atlas	Gotway (2002),[65] Haining (1994),[74] Rushton (2004),[75] Wang (2004),[43] Pukkala (1987)[76]
Protect privacy	Limit geographical detail Mask each individual location	McLaughlin (2002),[77] Krieger (2003),[19] Armstrong (1999),[78] French (2004),[13] Kwan (2004)[79]

See Appendix for definition of terms.

ual-level information to a variety of flexible, defined geographic areas, including not only census area units and municipalities or counties, but regions or areas defined by other sources of spatial data such as remote sensing imagery, in the case of environmental data. Such approaches will enable a foundation of data to be established from which other geocodes can be derived. Geocoding point coordinates, either as latitude, longitude, or UTM coordinates, provides such a foundational geocode and should be instituted as the basic unit. Along with the adoption of point-level geocodes for cancer cases, however, information on geocoding quality (metadata) should be linked to each individual case to ensure that geocoding error can be integrated into spatial analytic methods.[24] In addition, standards must be developed to ensure and maintain privacy and confidentiality of individual case information, to meet the goals of preserving and promoting public health while maintaining personal rights.

This publication was made possible through a Cooperative Agreement between the Centers for Disease Control and Prevention (CDC) and the Association of Schools of Public Health (ASPH), award number S-3111. Its contents are the responsibility of the authors and do not necessarily reflect the official view of the CDC or ASPH.

No financial conflict of interest was reported by the authors of this paper.

References

1. U.S. Bureau of the Census. Census Use Study: Health Information System II. Report No. 12. Washington, DC: U.S. Bureau of the Census, 1971.
2. U.S. Bureau of the Census. Census Use Study: The DIME Geocoding System. Report No. 4. Washington, DC: U.S. Bureau of the Census, 1970.
3. Centers for Disease Control and Prevention. Healthy people 2010, Vol. I. 2001. Available at: www.healthypeople.gov/document/html/volume2/23phi.htm. Accessed August 6, 2004.
4. Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. ZIP Code caveat: bias due to spatiotemporal mismatches between ZIP

Codes and U.S. census-defined geographic areas—the Public Health Disparities Geocoding Project. Am J Public Health 2002;92:1100–2.
5. Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health 2001;91:1114–6.
6. National Cancer Institute. Cancer Surveillance Research Implementation Plan. 1999. Available at: dccps.nci.nih.gov/DCCPS/SIG/. Accessed July 28, 2004.
7. Cromley EK, McLafferty SL. GIS and public health. New York: The Guilford Press, 2002.
8. Croner CM, Sperling J, Broome FR. Geographic information systems (GIS): new perspectives in understanding human health and environmental relationships. Stat Med 1996;15:1961–77.
9. U.S. Geological Survey. Geographic Names Information System (GNIS). Reston, VA: U.S. Geological Survey, 2004. Available at: www.census.gov/geo/www/tiger/. Accessed August 5, 2004.
10. Howe HL. Geocoding NY state cancer registry. Am J Public Health 1986;76:1459–60.
11. Bellander T, Berglind N, Gustavsson P, Jonson T, Nyberg F, Pershagen G, et al. Using geographic information systems to assess individual historical exposure to air pollution from traffic and house heating in Stockholm. Environ Health Perspect 2001;109:633–9.
12. Betts KS. Mapping the environment. Environ Health Perspect 1997;105:594–6.
13. French JL, Wand MP. Generalized additive models for cancer mapping with incomplete covariates. Biostatistics 2004;5:177–91.
14. Krautheim KR, Aldrich TE. Geographic information system (GIS) studies of cancer around NPL sites. Toxicol Ind Health 1997;13:357–62.
15. Lewis-Michl EL, Melius JM, Kallenbach LR, Ju CL, Talbot TO, Orr MF, et al. Breast cancer risk and residence near industry or traffic in Nassau and Suffolk Counties, Long Island, New York. Arch Environ Health 1996;51:255–65.
16. McLaughlin JR, Clarke EA, Nishri ED, Anderson TW. Childhood leukemia in the vicinity of Canadian nuclear facilities. Cancer Causes Control 1993;4:51–8.
17. Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Smith DF. Childhood cancer incidence rates and hazardous air pollutants in California: an exploratory analysis. Environ Health Perspect 2003;111:663–8.
18. Bach PB, Hoangmai H, Schrag D, Tate RC, Hargraves JL. Primary care physicians who treat blacks and whites. N Engl J Med 2004;351:575–84.
19. Krieger N, Zierler S, Hogan JW, Waterman P, Chen JT, Lemieux K, et al. Geocoding and measurement of neighborhood socioeconomic position: a U.S. perspective. In: Kawachi I, Berkman LF, eds. Neighborhoods and health. Oxford: Oxford University Press, 2003:147–78.
20. Broome FR, Meixler DB. The TIGER database structure. Cartogr Geogr Inf Sys 1990;17:39–47.
21. Marx RW. The TIGER system: yesterday, today and tomorrow. Cartogr Geogr Inf Sys 1990;17:89–97.
22. Cayo MR, Talbot TO. Positional error in automated geocoding of residential addresses. Int J Health Geogr 2003;2:10.
23. U.S. Bureau of the Census. TIGER®, TIGER/Line®, and TIGER-related products. Washington, DC: U.S. Bureau of the Census, 2004. Available at: www.census.gov/geo/www/tiger/. Accessed August 5, 2004.
24. Wiggins L, ed. Using geographic information systems technology in the collection, analysis, and presentation of cancer registry data: a handbook of basic practices. Springfield, IL: North American Association of Central Cancer Registries, 2002.
25. U.S. Postal Service. Acronyms and abbreviations. Washington, DC: U.S. Postal Service, 2004. Available at: www.usps.com/ncsc/lookups/usps_abbreviations.htm. Accessed August 6, 2004.
26. U.S. Postal Service. Address quality. Washington, DC: U.S. Postal Service, 2004. Available at: www.usps.com/ncsc/lookups/. Accessed August 6, 2004.
27. U.S. Postal Service. Coding accuracy support system. Technical guide 2004-2005 cycle. Washington, DC: U.S. Postal Service, 2004. Available at: www.ribbs.usps.gov/files/cass/casstech.pdf. Accessed August 6, 2004.
28. Yakich VR. Decoding addresses from point layers. ArcUser 2003;January–March:32–4.
29. Hurley SE, Saunders TM, Nivas R, Hertz A, Reynolds P. Post office box addresses: a challenge for geographic information system-based studies. Epidemiology 2003;14:386–91.
30. Yang DH, Bilaver LM, Hayes O, Goerge R. Improving geocoding practices: evaluation of geocoding tools. J Med Syst 2004;28:361–70.
31. U.S. Bureau of the Census. Cartographic boundary files. Washington, DC: U.S. Bureau of the Census, 2004. Available at: www.census.gov/geo/www/cob/scale.html. Accessed August 6, 2004.
32. Boscoe FP, Kielb CL, Schymura MJ, Bolani TM. Assessing and improving census tract completeness. J Regist Manage 2002;29:117–20.
33. McElroy JA, Remington PL, Trentham-Dietz A, Robert SA, Newcomb PA. Geocoding addresses from a large population-based study: lessons learned. Epidemiology 2003;14:399–407.
34. Krieger N, Chen JT, Waterman PD, Rehkopf DH, Subramanian SV. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—the public health disparities geocoding project. Am J Public Health 2003;93:1655–71.
35. Ratcliffe JH. Geocoding crime and a first estimate of a minimum acceptable hit rate. Int J Geogr Inf Sci 2004;18:61–72.
36. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology 2003;14:408–12.
37. Dearwent SM, Jacobs RR, Halbert JB. Locational uncertainty in georeferencing public health datasets. J Expo Anal Environ Epidemiol 2001;11:329–34.
38. Gregorio DI, Cromley E, Mrozinski R, Walsh SJ. Subject loss in spatial analysis of breast cancer. Health Place 1999;5:173–7.
39. Wing P, Reynolds C. The availability of physician services: a geographic analysis. Health Serv Res 1988;23:649–67.
40. Van Sickle J. Basic GIS coordinates. Boca Raton, FL: CRC Press, 2004.
41. Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial epidemiology: methods and applications. Oxford: Oxford University Press, 2000.
42. Krieger N, Chen JT, Waterman PD, Soobader MJ, Subramanian SV, Carson R. Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (U.S.). J Epidemiol Community Health 2003;57:186–99.
43. Wang F. Spatial clusters of cancers in Illinois 1986-2000. J Med Syst 2004;28:237–54.
44. U.S. Bureau of the Census. Census 2000 ZCTAs ZIP Code Tabulation Areas technical documentation. Washington, DC: U.S. Bureau of the Census, 2004. Available at: www.census.gov/geo/ZCTA/zcta_tech_doc.pdf. Accessed July 30, 2004.
45. Geronimus AT, Bound J, Neidert LJ. On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. J Am Stat Assoc 1996;91:529–37.
46. Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA. LandScan: A global population database for estimating populations at risk. Photogrammetric Engineering & Remote Sensing 2000;66:849–57.
47. Fisher PF, Langford M. Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. Environ Plan A 1995;27:211–24.
48. Flowerdew R, Green M. Areal interpolation and types of data. In: Fotheringham AS, Rogerson P, eds. Spatial analysis and GIS. London: Taylor and Francis, 1994:121–45.
49. Tobler WR. Smooth pycnophylactic interpolation for geographical regions. J Am Stat Assoc 1979;74:519–30.
50. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. Environ Health Perspect 1997;105:598–605.
51. Cromley EK. GIS and disease. Annu Rev Public Health 2003;24:7–24.
52. Gunier RB, Harnly ME, Reynolds P, Hertz A, Von Behren J. Agricultural pesticide use in California: pesticide prioritization, use densities, and population distributions for a childhood cancer study. Environ Health Perspect 2001;109:1071–8.
53. Brimberg J, Love R. A new distance function for modeling travel distances in a transportation network. Transport Sci 1992;26:129–37.
54. Love D, Lindquist P. The geographical accessibility of hospitals to the aged: a geographic information systems analysis within Illinois. Health Serv Res 1995;29:629–51.
55. Rushton G. Methods to evaluate geographic access to health services. J Public Health Manag Pract 1999;5:93–100.
56. Williams AP, Schwartz WB, Newhouse JP, Bennett BW. How many miles to the doctor? N Engl J Med 1983;309:958–63.
57. Phibbs C, Luft H. Correlation of travel time on roads versus straight line distance. Med Care Res Rev 1995;52:532–42.
58. Francis RL, Lowe TJ, Rushton G, Rayco MB. A synthesis of aggregation methods for multifacility location problems: strategies for containing error. Geogr Anal 1999;31:67–87.

59. Hewko J, Smoyer-Tomic KE, Hodgson MJ. Measuring neighbourhood spatial accessibility to urban amenities: does aggregation error matter? Environ Plan A 2002;34:1185–206.
60. Lovett A, Haynes R, Sunnenberg G, Gale S. Car travel time and accessibility by bus to general practitioner services: a study using patient registers and GIS. Soc Sci Med 2002;55:97–111.
61. Penchansky R, Thomas JW. The concept of access: definition and relationship to consumer satisfaction. Med Care 1981;19:127–40.
62. Fortney J, Rost K, Warren J. Comparing alternative methods of measuring geographic access to health services. Health Serv Outcomes Res Methodol 2000;1:173–84.
63. Miller HJ. Measuring space-time accessibility benefits within transportation networks: Basic theory and computational procedures. Geogr Anal 1999;31:187–212.
64. Miller HJ, Wu Y-H. GIS software for measuring space-time accessibility in transportation planning and analysis. GeoInformatica 2000;4:141–59.
65. Gotway CA, Young LJ. Combining incompatible spatial data. J Am Stat Assoc 2002;97:632–48.
66. Flowerdew R, Amrhein C. Poisson regression models of Canadian census division migration flows. Pap Reg Sci Assoc 1989;67:89–102.
67. Flowerdew R, Green M. Data integration: statistical methods for transferring data between zonal systems. In: Masser I, Blakemore M, eds. Handling geographical information: methodology and potential applications. Harlow Essex, UK: Longman Publishing Group, 1991:38–54.
68. Markoff J, Shapiro G. The linkage of data describing overlapping geographical units. Hist Methods Newsl 1973;7:34–46.
69. Goodchild MF, Lam NS-N. Areal interpolation: a variant of the traditional spatial problem. Geo-Process 1980;1:297–312.
70. Lam NS. Spatial interpolation methods: a review. Am Cartogr 1983;10:129–49.
71. Goodchild MF, Anselin L, Deichmann U. A framework for the areal interpolation of socioeconomic data. Environ Plan A 1993;25:383–97.
72. Sadahiro Y. Accuracy of areal interpolation: a comparison of alternative methods. J Geogr Syst 1999;1:323–46.
73. Langford M, Maguire DJ, Unwin DJ. The areal interpolation problem: estimating population using remote sensing in a GIS framework. In: Masser I, Blakemore M, eds. Harlow Essex, UK: Longman Publishing Group, 1991:55–77.
74. Haining R, Wises S, Blake M. Constructing regions for small area analysis: material deprivation and colorectal cancer. J Public Health Med 1994;16:429–38.
75. Rushton G, Peleg I, Banerjee A, Smith G, West M. Analyzing geographic patterns of disease incidence: rates of late-stage colorectal cancer in Iowa. J Med Syst 2004;28:223–36.
76. Pukkala E, Gustavson N, Teppo L. Atlas of cancer incidence in Finland 1953-82. Cancer Soc Finland Pub No. 37. Helsinki: Finnish Cancer Registry, 1987.
77. McLaughlin CC. Confidentiality protection in publicly released central cancer registry data. J Regist Manage 2002;29:84–8.
78. Armstrong MP, Rushton G, Zimmerman DL. Geographically masking health data to preserve confidentiality. Stat Med 1999;18:497–525.
79. Kwan M-P, Casas I, Schmitz BC. Protection of geoprivacy and accuracy of spatial information: how effective are geographical masks? Cartographica 2004;39:15–28.
80. Cate FH. Privacy in the information age. Washington, DC: Brookings Institute, 1997.
81. Katsh ME. Law in a digital world. New York: Oxford University Press, 1995.
82. Solove DJ, Rothenberg M. Information privacy law. New York: Aspen Publishers, 2003.
83. Armstrong MP. Geographic information technologies and their potentially erosive effects on personal privacy. Stud Soc Sci 2002;27:19–28.
84. Curry MR. The digital individual and the private realm. Ann Assoc Am Geogr 1997;87:681–99.
85. Dobson J. Is GIS a privacy threat? GeoWorld 1998;11(7).
86. Dobson J, Fisher P. Geoslavery. IEEE Technol Society 2003;Spring:47–52.
87. Goss J. We know who you are and we know where you live: the instrumental rationality of geodemographic systems. Econ Geogr 1995;71:171–98.
88. Monmonier M. Spying with maps: surveillance technologies and the future of privacy. Chicago, IL: University of Chicago Press, 2002.
89. Waters N. GIS and the bitter fruit: privacy issues in the age of the internet. GeoWorld [serial on the Internet], 2000. Available at: www.geoplace.com/gw/2000/0500/0500edg.asp. Accessed August 6, 2005.
90. Gostin L. Health information privacy. Cornell Law Rev 1995;80:451–528.
91. Health Privacy Working Group Health Privacy Project. Principles for health privacy. Washington, DC: Institute For Health Research and Policy, Georgetown University, 1999.
92. Pritts J. The state of health privacy: an uneven terrain, a comprehensive survey of state health privacy statutes. Washington, DC: Health Privacy Project, Institute for Health Care Research and Policy, Georgetown University, 1999.
93. Roarch WH. Medical records and the law. Gaithersburg, MD: Aspen Publishers, 1998.
94. Hodge JG, Gostin LO. Public health practice vs. research, a report for public health practitioners including cases and guidance for making distinctions. Baltimore, MD: Johns Hopkins Bloomberg School of Public Health, Center for Law and Public's Health, 2004.
95. U.S. Dept. Health Human Services. Standards for privacy of individually identifiable health information. 45 CFR Parts 160, 164, 2002.
96. Centers for Disease Control and Prevention. HIPAA privacy rule and public health, guidance from CDC and the U.S. Department of Health and Human Services. MMWR Morbi Mortal Wkly Rep 2003;52(suppl)17, 19–20.
97. Gittler J. State cancer registries: compendium of state privacy and security statutes and administrative regulations. Iowa City, Iowa: National Health Law & Policy Resource Center, College of Law, University of Iowa, 2004.
98. North American Association of Central Cancer Registries. NAACCR workshop report: data security and confidentiality. Springfield, IL: North American Association of Central Cancer Registries, 2002.
99. European Network of Cancer Registries. Guidelines on confidentiality in population-based cancer registration in the European Union. Lyons: IARC, 2001.
100. Diggle PJ. Point process modelling in epidemiology. In: Barnett V, Turkman KF, eds. Statistics for the environment. New York: Wiley, 1993:89–110.
101. Waller LA. Statistical power and design of focused clustering studies. Stat Med 1996;15:765–82.
102. Arbia G, Griffith D, Haining R. Error propagation modelling in raster GIS: overlay operations. Int J Geogr Inf Sci 1998;12:145–67.
103. Cressie N, Kornak J. Spatial statistics in the presence of location error with an application to remote sensing of the environment. Stat Sci 2003;18:436–56.
104. Gabrosek J, Cressie N. The effect on attribute prediction of location uncertainty in spatial data. Geogr Anal 2002;34:262–85.
105. Chakraborty J, Armstrong MP. Assessing the impact of airborne toxic releases on populations with special needs. Prof Geogr 2001;53:119–31.
106. LeTendre D, Cress RD, Riddle S, Creech CM. Where did they really live? Resolving discrepancies in address at diagnosis. J Regist Manage 2000;27:57–8.
107. Federal Geographic Data Committee. Address data content standard: public review draft, 2003. Available at: www.census.gov/geo/www/standards/scdd/AddressStandardV2_April%2017_2003.pdf. Accessed August 4, 2004.
108. Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel JF, Bertollini R. Disease mapping and risk assessment for public health. New York: John Wiley & Sons, 1999.

Appendix

Glossary of terms used: E911, "Enhanced 911" (a technology that enables emergency services to locate the geographic position of the caller); FIPS, Federal Information Processing Standard; HRSA, Health Resources Services Administration; HIPAA, Health Insurance Portability and Accountability Act of 1996 (Title II addresses the security and privacy of health data); NAACCR, North American Association of Central Cancer Registries; PO Box, post office box number; PCSA, primary care service areas (developed from ZIP code numbers on a large sample of Medicare claims data); Spatial Metadata, data that describes the characteristics and sources of spatial data; SoundEx, a group of algorithms that allow coding of phonemes and substitutions of phonetically similar character strings to enhance database searches for variant spellings and misspellings, resulting in increased error-tolerance; SoundEx algorithms can be used to enhance database matching of literal strings, for example of names of persons and streets; ZCTA: ZIP code tabulation areas (aggregates of census blocks that most closely correspond with ZIP code areas).